Szemantikus szerepek automatikus címkézése függőségi elemző alkalmazásával magyar nyelvű gazdasági szövegeken

Subecz Zoltán

Szolnoki Főiskola
5000 Szolnok, Tiszaligeti sétány 14.
subecz@szolf.hu

Kivonat: Jelen tanulmányunkban bemutatjuk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, amely automatikusan képes magyar nyelvű szövegekben szemantikus szerepek címkézésére függőségi elemző alkalmazásával. Munkánkban a vállalati vásárlások, tulajdonváltozások keretével foglalkoztunk. Jellemzőkészletünkben felszíni, morfológiai és a függőségi elemzés alapján kinyert jellemzőket használtunk fel. Ezen alapjellemzőket kiegészítettük a jellemzőkből számolt statisztikai arányokkal is. Megvizsgáltuk, hogy a modell hogyan teljesít egy gyakori célszóra önállóan, és a célszavak keretekbe összefoglalt csoportjára is.

1 Bevezetés

Az Információkinyerés egyik fontos feladata a névelemek felismerése mellett az események detektálása [15,16]. A szövegekben lévő események felismerése, analizá-lása, és hogy hogyan viszonyulnak egymáshoz időben, fontos a szöveg tartalmának megismerésében. Az események detektálása mellett fontos azok szemantikus kapcsolatainak, vagy szemantikus szerepeinek megtalálása is (szemantikus szerepek címkézése, Semantic Role Labeling, SRL). Az események és azok szemantikus szerepeinek detektálását a természetes nyelvfeldolgozás sok területén lehet hasznosítani. Például az összegzéskészítés, gépi fordítás és a válaszkeresés területein.

Munkánkban a *szemantikus szerepek címkézésével* foglalkoztunk. Ez a szemantikus kapcsolatok azonosítását jelenti egy *szemantikus kereten* belül (semantic frame). A *keretek* eseményeket írnak le azok szereplőinek szintaktikai és szemantikai megkötésein keresztül. Munkánkban a *vállalati vásárlások, tulajdonváltozások* keretével foglalkoztunk.

A *szemantikus szerepek címkézése* napjainkban a természetes nyelvfeldolgozás (NLP) egyik legdinamikusabban fejlődő területe. Angol nyelvű szövegekre általában *konstituensfa alapú* szintaktikai elemzőt használnak az elő-feldolgozásnál az angol nyelv erősen konfiguratív tulajdonsága miatt, ahol is a legtöbb mondatszintű szintaktikai információt a szórenddel fejeznek ki. Ezzel ellentétben a magyar nyelv gazdag morfológiával és szabad szórenddel rendelkezik. A *függőségi fákkal* dolgozó elemzők különösen jól használhatóak szabad szórendű nyelvek elemzésére, így a magyarra is,

ezek ugyanis könnyebben teszik lehetővé az egymással nem szomszédos, de összetartozó szavak összekapcsolását is. Ezért mi a magyar nyelvű szövegeinkre *függőségi fákkal dolgozó elemzőt* használtunk a *magyarlanc* programcsomag segítségével [20]. A szövegek szavakra bontására, a szavak morfológiai elemzésére, szófaji egyértelműsítésére, és mondatok függőségi nyelvtan szerinti szintaktikai elemzésére is ezt alkalmaztuk.

A *szerepek* a legegyszerűbb esetekben a célszó *szintaktikai kapcsolatai* voltak, de nem mindig. Sokszor a keresett szerep távol helyezkedett el a függőségi fában a célszótól, gyakran a mondat másik felében. És olyan is volt, hogy a szintaktikai kapcsolat alapján várt helyen nem a keresett szerep volt. Ez utóbbi gyakran a szintaktikai elemző hibájából adódott. Így a feladat a függőségi fában a célszótól távolabbi szerepek megkeresése és a közelebbi hamis pozitív jelöltek kiszűrése volt.

2 Kapcsolódó munkák

A *szemantikus szerepek címkézése* napjainkban a természetes nyelvfeldolgozás (NLP) egyik legdinamikusabban fejlődő területe. Angol nyelvű szövegekre sok módszer született már, ezek általában konstituensfa alapú szintaktikailag elemzett mondatokat használnak, és mondat szinten vizsgálják az eseményeket.

Kezdetben az SRL munkákban csak igékkel foglalkoztak, az igéket önállóan vizsgálták és általános szerepeket kerestek (például Agent, Patient, Instrument). A PropBank korpusz [13] szövegeit használták fel, amiben angol nyelvű szövegekhez vannak kiemelt igék annotálva a hozzájuk tartozó szemantikus szerepekkel. A 2004-es és 2005-ös CoNLL feladatokban foglalkoztak ezzel a témával [2,4].

Később az igéket már nem önállóan vizsgálták, hanem tématerületenként csoportosították azokat (keretek). Az általános szerepek mellett már vizsgáltak doménspecifikus szerepeket is. Ehhez a FrameNet korpusz [7] szövegeit használták fel, amiben angol nyelvű szövegek vannak szemantikus szerepek szerint annotálva. Ezek is elsősorban igékkel foglalkoznak, de keresnek nem igei célszavakra is. Egy fontos alaptanulmányt készített D. Gildea és D. Jurafsky [8] az SRL témában. A Senseval-3 task [11] egyik része is a FrameNet-re alapozott SRL feladat volt. Az ACE program is más NLP feladatok mellett SRL témával is foglalkozik [1].

Xue és társa [19] a jelöltek számának csökkentésére mutatott be egy módszert. A jelöltek számát jelentősen csökkentették, miközben a fedést magasan tartották.

Koomen és társai [10] és Toutanova és társai [18] a szerepek azonosítása után a szerepek közötti kapcsolatokkal, függőségekkel foglakoztak. Azt vizsgálták, hogy a megtalált kifejezések hogyan lehetnek együtt a célszónak a szerepei.

Surdeanu és társai [17] és Pradhan és társai [14] számos SRL alapú rendszer kimenetét kombinálták egy rendszerbe.

Carreras és társai [2,3] és Surdeanu és társai [17] munkáiban, ha egy mondaton belül több célszó található, akkor ezeket nem csak egymástól függetlenül kezelték, hanem közös szerepeket is kerestek hozzájuk.

Johansson és társa [9] angol nyelvű szövegekre a konstituens alapú elemzés helyett függőségi elemzést használt.

Szemantikus szerepek címkézésére magyar nyelvű szövegekre is készültek már munkák. Farkas és társai [3] a szemantikuskeret-illesztésre *szabály alapú* módszert használtak. Mi gépi tanulásos módszert alkalmaztunk ugyanerre. A szabályalapú módszerrel ellentétben a gépi tanulásos módszer nem igényel annyi erőforrást és előfeldolgozást, és automatikusan alkalmazható más doménekre is. Ehmann és társai [4] pszichológiai témájú szövegeken szemantikus szerepek címkézésénél csak két általános szerepet keresnek: az ágens és a recipiens szerepeket (cselekvő, elszenvedő). Mi a vállalatfelvásárlások kereten belül nem csak a két általános szerepet, hanem több domén-specifikus szerepet is címkéztünk. A következő szerepeket vizsgáltuk: Vevő, Eladó, Áru, Ár, Idő. Csak az igei és főnévi igenévi célszavak szerepeit kerestük.

Az angol nyelvű szövegekre általában *konstituensfa* alapú szintaktikailag elemzett mondatokat használnak. Az előző pontban ismertetett okok miatt mi a magyar nyelvű szövegeinkre *függőségi fákkal dolgozó elemzőt* használtunk a *magyarlanc* programcsomag segítségével [20].

3 Szemantikus keretek és a szemantikus szerepek

Sok információkinyerő rendszer manapság *tárgykör (domén)* specifikus *keretekkel* dolgozik. Egy-egy tárgykör eseményeit célszerű egy *kereten* belül vizsgálni, hiszen ugyanazok a *szerepek* tartoznak minden eseményhez, ami egy adott csoporthoz tartozik. Például egy repülőjegy foglalásokat feldolgozó rendszer a következő *szerepeket* használhatja: indulási időpont, érkezési időpont, célállomás, indulási állomás, távolság, ár. Az előző rendszer *célszavai* lehetnek például: foglal, lefoglal, előjegyez, vált. Ha a célszavakat önállóan dolgozzuk fel, akkor csak kevés tanító adattal tudunk dolgozni. A célszavak *keretekben történő csoportosítása* jelentősen csökkenti ezt a problémát, hiszen a több célszó tanító adatai összeadódnak.

Munkánkban a *vállalati vásárlások, tulajdonváltozások keretével* foglalkoztunk, *igei és főnévi igenévi* célszavakhoz kerestük ki a szereplőket. A következő igei célszavakat vizsgáltuk meg az adott kereten belül: *vesz, vásárol, szerez, bekebelez, gyarapít, ad, átruház, értékesít, forgalmaz*. Valamint e célszavak minden igekötős, módbeli és időbeli változatát is. A célszavakhoz a mondatokon belül a következő szerepeket kerestük meg: *vevő, eladó*, *áru, ár, idő*.

Példák a szerepekre a vállalati vásárlások tárgykörben. A példákban vastag betűvel vannak kiemelve a *célszavak* és szögletes zárójelben a *szerepek* találhatóak. Alsóindexben szerepel az adott szerep típusa.

1. *[A svéd Electrolux]$_{Eladó}$ **eladja** [motorgyártó részlegét]$_{Áru}$ [az olasz Appliance Components Companies részvénytársaságnak]$_{Vevő}$ – tájékoztatott az Electrolux.*
2. *[A Deutsche Börse AG]$_{Vevő}$ pénteken bejelentette, hogy teljesen **megveszi** [a luxemburgi Clearstream elszámolóházat]$_{Áru}$.*
3. *[A Royal Dutch Shell csoport]$_{Vevő}$ [400 millió dollárért]$_{Ár}$ **megvenni** készül [a legnagyobb kínai offshore-földgáz- és olajmező 20 százalékát]$_{Áru}$.*
4. *[A svéd Ericsson]$_{Eladó}$ bejelentette, hogy [a német Infineonnak]$_{Vevő}$ **adja** el [chipgyártó részlegét]$_{Áru}$, [400 millió euróért]$_{Ár}$.*

5. *[A többnyire szárazföldi szállítással foglalkozó magyar tulajdonban lévő Cronus Kft.]$_{Vevő}$ [a közelmúltban]$_{Idő}$* ***megvásárolta*** *[a Magyar Államvasutak Rt.-től]$_{Eladó}$ [a debreceni székhelyű MÁV Hajdú Vasútépítő-mélyépítő Kft.-t]$_{Áru}$ – jelentette be szerdán Debrecenben a Cronus Kft. tulajdonosa.*

A példákban látszik, hogy egy szerep általában *több szóból* áll és a mondatok általában nem tartalmazzák mind az öt szerepet.

3.1 Felhasznált Korpusz

Az alkalmazásunk teszteléséhez a *Szeged Korpusznak* a rövidhírek csoportjának egy olyan változatát használtuk fel, amelyikben annotálva vannak a vállalati vásárlásokra a szemantikus szerepek. Ezek közül 1000 mondatot használtunk fel. A tanításhoz és kiértékeléshez 10-szeres keresztvalidációt alkalmaztunk.

3.2 Felhasznált programcsomagok

A feladatokat *bináris osztályozásra* vezettük vissza. Az osztályozáshoz a *Weka* programcsomagnak[1] a J48-as döntési fa elemzőjét használtuk fel. A Weka adatbányászati feladatokhoz készített gépi tanuló algoritmusok gyűjteménye. A feladathoz felhasználtuk még a magyarlanc 2.0 programcsomagot is. [20] A csomag magyar szövegek mondatra és szavakra bontására, a szavak morfológiai elemzésére, majd szófaji egyértelműsítésére, és mondatok függőségi nyelvtan szerinti szintaktikai elemzésére alkalmazható.

4 A magyarlanc programcsomag elemzésének bemutatása

A magyarlanc a bemenetére érkező mondatoknak elkészíti az előző pontban leírt elemzését. A mondat minden szavához külön sorba elkészíti az elemzést (1. ábra). Minden szóról megadja a következő információkat: *sorszám, szó, lemma, szófaj, morfológiai kódok*. A sor végén megadja, hogy az adott szó melyik szóval van *szintaktikai kapcsolatban*, és hogy milyen a *kapcsolat típusa*. A szintaktikai kapcsolatok alapján a mondatok egy *függőségi fát* alkotnak.

Az elemzés után megjelenítettük *vizuális elemzővel* a szintaktikai kapcsolatok alapján a mondat függőségi fáját a program online elemzőjével[2] (2. ábra). Az elemzés és a vizuális ábrázolás egymásnak megfelelően megadja a szavak közötti *szintaktikai kapcsolatokat*.

Példa:

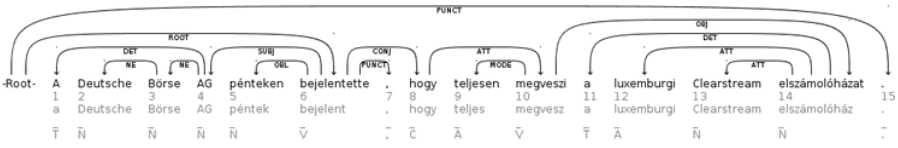
A Deutsche Börse AG pénteken bejelentette, hogy teljesen megveszi a luxemburgi Clearstream elszámolóházat.

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://www.inf.u-szeged.hu/rgai/magyarlanc-service/

```
1       A       a       T       SubPOS=f        4       DET
2       Deutsche        Deutsche        N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     3       NE
3       Börse   Börse   N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     4       NE
4       AG      AG      N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     6       SUBJ
5       pénteken        péntek  N       SubPOS=c|Num=s|Cas=p|NumP=none|PerP=none|NumPd=none     6       OBL
6       bejelentette    bejelent        V       SubPOS=m|Mood=i|Tense=s|Per=3|Num=s|Def=y       0       ROOT
7       ,       ,       ,       _       6       PUNCT
8       hogy    hogy    C       SubPOS=s|Form=s|Coord=p 6       CONJ
9       teljesen        teljes  A       SubPOS=f|Deg=p|Num=s|Cas=w|NumP=none|PerP=none|NumPd=none       10      MODE
10      megveszi        megvesz V       SubPOS=m|Mood=i|Tense=p|Per=3|Num=s|Def=y       8       ATT
11      a       a       T       SubPOS=f        14      DET
12      luxemburgi      luxemburgi      A       SubPOS=f|Deg=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none       14      ATT
13      Clearstream     Clearstream     N       SubPOS=p|Num=s|Cas=n|NumP=none|PerP=none|NumPd=none     14      ATT
14      elszámolóházat  elszámolóház    N       SubPOS=c|Num=s|Cas=a|NumP=none|PerP=none|NumPd=none     10      OBJ
15      .       .       .       _       0       PUNCT
```

1. ábra.



2. ábra.

Az elemzésekből látszik, hogy a függőségi elemző egy *szabályos elemző fát* készít. A fa legfelső eleme a *Root*. A fa *csomópontjaiban* vannak a mondat szavai, az *ágak* a szavak közötti *szintaktikai kapcsolatokat* reprezentálják. A fában kiemelt szerepe van az *igéknek*. A *főige* (a példákban a *bejelentette*) általában a Root alatt helyezkedik el, a szintaktikai kapcsolatokon keresztül ehhez kapcsolódnak a többi elemek.

Ha a szerep *több szóból* áll, akkor ezek a szavak egy *részfát* alkotnak a mondat fáján belül. A részfa a *kiemelt szaván* (fejszó, headword) keresztül kapcsolódik a fa többi részéhez.

Van, amikor a szerep kiemelt szava (headword) a célszóhoz kapcsolódik közvetlenül. Ilyen esetben könnyebb megtalálni a szerepet. Van, amikor a szerep kiemelt szava nem a célszóhoz kapcsolódik közvetlenül. A példamondatnál a vevő kiemelt szava (AG) nem kapcsolódik szintaktikailag a *megveszi* célszóhoz az elemzőfában, hanem a *bejelentette* igén keresztül. Ilyenkor nehezebb megtalálni a szerepet. Minél közelebb van a szerep a célszótól a mondaton vagy az elemzőfán belül, annál nagyobb a valószínűsége a szerep azonosításának.

Bár a magyarlanc program elkészíti a mondatoknak a szintaktikai elemzését, de a példákon is láttuk, hogy a szintaktikai kapcsolat típusából nem következik a szemantikai szerep. Például a *vesz* célszónak az alanya a vevő, az *elad* célszónak az alanya az eladó. Így a szintaktikai kapcsolatok mellett több más tulajdonságot is meg kell figyelni a mondatban. A feladatot megnehezíti, hogy a magyarlanc elemző is hibával dolgozik, így ez a hiba és a hibákból eredő hamis döntések megjelennek a mi eredményeinkben is. Jobb eredményeket kaptunk volna, ha szövegeink kézzel lettek volna annotálva ezen szempontok szerint.

5 Az osztályozás bemutatása

A célszavakhoz a következő szerepeket vizsgáltuk: *vevő, eladó, áru, ár, idő*. Minden bemeneti mondatnál adott volt a *célszó*. A feladat az adott szerep megkeresése volt.

Az osztályozóknál a *jelöltek* a függőségi elemzőfa csomópontjai voltak. Egy mondaton belül általában egy csomópont a keresett szerep kiemelt szava (headword). Az osztályozásnál ezek a *true* esetek, a többi csomópont pedig a *false* eset.

Az osztályozáshoz *bináris osztályozót* használtunk. Az osztályozó az adott mondatnál bejelöli a keresett szerepet. Az osztályozónak *nem adtuk meg*, hogy az adott mondat tartalmazza-e az adott szerepet, vagy sem. Voltak olyan mondatok is, amelyek nem tartalmazták a keresett szerepet. (1. táblázat)

A kiértékelésnél *szigorú szabályt* alkalmaztunk: csak azt a döntést fogadtuk el, amelyik pontosan az annotált szerepet jelöli meg. Sem az ezt tartalmazó fákat, sem ennek a részfáit nem fogadtuk el pozitív döntésnek. Ha ennél enyhébb szabályt alkalmaznánk, akkor magasabb eredményeket kapnánk.

5.1 Jellemzőkészlet

A tanító és a kiértékelő halmazon a *jelöltekhez jellemzőket* vettünk fel. Az SRL feladatokban használt általános jellemzőket [8] mi is alkalmaztuk. Ezeken kívül újakkal is kibővítettük a jellemzőkészletünket. Ehhez felhasználtuk a *függőségi elemzőfát* is, a jelölt és a célszó viszonyát a függőségi fában, mert ez gyakran egy fontos tulajdonsága az adott szerepnek.

A jelöltekhez a *következő jellemzőket* választottuk ki:

Felszíni jellemzők: *Bigramok, trigramok*: A vizsgált szavak végén lévő 2-es, 3-as betűcsoportok. *Pozíció*: a jelölt a célszó előtt vagy után áll a mondatban. *Távolság-mondatban*: a jelölt és a célszó szótávolsága a mondaton belül.

Morfológiai jellemzők: Mivel a magyar nyelv igen gazdag morfológiával rendelkezik, ezért számos morfológiaalapú jellemzőt definiáltunk. Jellemzőként definiáltuk az eseményjelöltek MSD-kódját felhasználva a következő morfológiai jegyeket: *típus*(SubPos), *mód*(Mood), *eset*(Cas), *idő*(Tense), *személy*(PerP), *szám*(Num), *határozottság*(Def). *Szófaj, Lemma*: a jelölt és a célszó szófaja és lemmája.

Jellemzők az elemzőfa alapján-1: Ide azokat a jellemzőket soroltuk, amelyeket az SRL feladatokhoz általában felhasználnak [8]. A jelölt és a célszó viszonyát vizsgáltuk a függőségi elemzőfában. Mindkettő egy-egy csomópont az elemzőfában. *Szófaj-útvonal*: Egymás után írtuk a jelölt és a célszó közötti csomópontok szófaját, feljegyezve azt is, hogy az elemzőfában felfelé, vagy lefelé haladtunk az adott kapcsolatnál. Például: C↑S↑V↑C↑V↑V↓V↓N↓N↓A. *Uralkodó-kategória-szófaja*: A jelölt és a célszó közötti útvonalon megkerestük a legmagasabban fekvő csomópontot, és feljegyeztük a hozzá tartozó szó szófaját.

Jellemzők az elemzőfa alapján-2: Itt az egyéni, új jellemzőket soroltuk fel. *Jelölt-célszó-távolság-elemzőfában*: A jelölt és a célszó csomópontjai közötti csomópontok száma az elemzőfában. *Lemma-útvonal*: Mint a Szófaj-útvonal, de itt a jelölt és a célszó között végigmenve a csomóponti szavak lemmáját jegyeztük fel. Például: Budapesti↑Értéktőzsde↑honlap↑közöl↓megvásárol. *Szintaktikai-kapcsolat-útvonal*: Az

előzőhöz hasonlóan itt azt vettük fel, hogy a jelölt és a célszó között az elemzőfában milyen szintaktikai kapcsolatokon keresztül tudunk eljutni. Például: ↑COORD*SUBJ↓ATT↓INF↓OBJ↓ATT. *Jelölt-alatti-részfában-van-e-névelem*: A magyarlanc program az elemzésében jelöli, ha talált névelemeket a mondatban. Mivel a vállalati tulajdonváltozások témakörében gyakran találkozunk vállalati névelemekkel, ezért felvettük, hogy a jelölt, vagy az alatta levő részfa tartalmaz-e névelemet? *Jelölt-alatti-részfában-névelem-távolság*: az előzőhöz hasonlóan megadtuk a részfában azt a mélységet, ahol először találtunk névelemet.

5.2 Statisztikai arány felhasználása az osztályozásnál

A jelöltekhez a jellemzőket *két módszer* alapján választottuk ki. *Első módszernél* az előző részben bemutatott alapjellemzőket használtuk fel. *Második módszernél* az alapjellemzők helyett a tanító adatokon a jellemzőkészletből számított statisztikai arányokat használtuk fel: a tanító halmaz alapján megszámoltuk minden jellemző esethez, hogy hány alkalommal fordult elő és ebből hányszor volt a jelölt *pozitív*. Ezek alapján kiszámítottuk a hozzá tartozó pozitív-arányt. Például ha a *Jelölt-lemma* jellemzőnél a *jelölt-lemma = Corp.* eset 11-szer fordult elő és ebből 7-szer volt *pozitív* eset (4-szer pedig negatív), akkor hozzá a 0,64-es pozitív-arány tartozott. Ebben az esetben az osztályozónak a jelöltekhez nem az alapjellemzőt, hanem a hozzá tartozó arányt adtuk meg. Az előző példánál *Jelölt-lemma-arány* = 0,64. Ezzel *jelentősen csökkentettük az osztályozó vektorterének méretét* az első módszerhez képest és így a futási időt is. Ez a kidolgozási időszakban hasznos volt. *Harmadik esetben* az előző két módszer jellemzőit együtt használtuk fel.

A statisztikai-arány jellemzők hatása az osztályozás eredményére. Megvizsgáltuk, hogy az előzőleg bemutatott *statisztika-arány jellemzők* hogyan befolyásolják az osztályozási eredményeinket. Először az osztályozást lefuttattuk csak a statisztikai-arány jellemzőkkel, majd csak az alapjellemzőkkel és végül a két jellemzőcsoporttal együtt. Azt tapasztaltuk, hogy az alapjellemzőkkel eset önállóan általában jobban teljesített, mint a statisztikai-arány eset önállóan. De a *legjobb eredményt* akkor kaptuk, amikor az alapjellemzőket és a statisztikai-arány jellemzőket együtt használtuk.

5.3 Vektortér méretének csökkentése

A *vektortér méretét csökkentettük* a következő módszerrel: csak azokat a *jellemző-előfordulásokat* vettük fel az osztályozáshoz, amelyek a tanító halmazon *legalább háromszor* szerepeltek. Ezzel *jelentősen csökkentettük a futási időt* és csak az osztályozás szempontjából jelentéktelen jellemző-előfordulásokat hagytuk ki.

5.4 Célszavak csoportosítása a kereten belül

Először a modell viselkedését egy gyakori célszóra önállóan néztük meg. Ehhez a *vásárol* célszót választottuk ki.

Majd a célszavakat csoportosítottuk. A vásárlásokkal kapcsolatos mondatoknál a *vevő* és az *eladó* szerepek viselkedését meghatározza, hogy az adott célszónál az alany általában vevő vagy eladó. Ezért a célszavakat két csoportra bontottuk a következő egyszerű módszerrel. A *vevő-centrikus* csoportba azok a szavak kerültek, amelyeknél az alany általában a *vevő*: vesz, vásárol, szerez, bekebelez, gyarapít. Az *eladó-centrikus* csoportba pedig azok, amelyiknél az alany általában az *eladó*: ad, átruház, értékesít, forgalmaz. Ez a felosztás segítette a *vevő* és az *eladó* szerepek megtalálását. Egy harmadik esetben pedig nem végeztünk csoportosítást.

5.5 Baseline mérések

A Baseline módszereket a *döntési fa legfontosabb feltételei alapján* állítottuk össze.

Azokat a jelölteket vettük pozitívnak, amelyekre teljesül:

Az *Áru szerepnél* azokat, amelyek tárgy (OBJ) szintaktikai kapcsolatban vannak a célszóval.

Az *Ár szerepnél* azokat, amelyeket egy előre elkészített pénznemek lista tartalmazott.

Az *Idő szerepnél* azokat, amelyeket a következő lista tartalmazott: évszámok 1990-2014-ig, hónapnevek, napnevek, sorszámok 1-31-ig.

A *vevő-centrikus célszavaknál a Vevő szerepnél* és az *eladó-centrikus célszavaknál az Eladó szerepnél* azokat, amelyek alany (SUBJ) kapcsolatban vannak a célszóval.

A *vevő-centrikus célszavaknál az Eladó szerepnél* azokat, amelyek végén a következő trigramok állnak: tól, től, ból, ből.

Az *eladó-centrikus célszavaknál a Vevő szerepnél* azokat, amelyek részes eset (DAT) kapcsolatban vannak a célszóval.

A következő eredményeken látni fogjuk, hogy gépi tanulási modell jóval *felülteljesítette a Baseline modellünket*.

5.6 Statisztikai adatok

Mondatok száma összesen: 1000 db. Azon mondatok száma, amelyek tartalmazzák az adott szerepet:

1. táblázat. Statisztikai adatok (db).

Célszavak	**Mondatok száma**	**Vevő**	**Eladó**	**Áru**	**Ár**	**Idő**
kiemelt: *vásárol*	265	263	107	276	104	99
Vevő-centrikus	548	531	222	573	214	208
Eladó-centrikus	452	261	374	459	82	115
csoportosítás nélkül	1000	783	579	1025	299	312

Az osztályozónak *nem adtuk meg*, hogy az adott mondat tartalmazza-e az adott szerepet, vagy sem. (Az Áru szerep azért nagyobb, mint a mondatok száma, mert volt olyan mondat, ahol több áru szerepelt.)

6 Eredmények

6.1 Baseline mérések eredményei

2. táblázat. Baseline mérések eredményei.

Szerep	**Pontosság**	**Fedés**	**F-mérték**
Vevő-centrikus célszavak			
Vevő	48,24	59,73	**53,37**
Eladó	54,77	72,13	**62,26**
Áru	73,25	73,25	**73,25**
Ár	67,33	96,02	**79,16**
Idő	34,74	57,89	**43,42**
Eladó-centrikus célszavak			
Vevő	78,18	44,10	**56,39**
Eladó	42,63	47,50	**44,93**
Áru	77,47	72,97	**75,15**
Ár	62,64	93,44	**75,00**
Idő	23,95	46,51	**31,62**

6.2 Eredmények a *vásárol* kiemelt célszóra

3. táblázat. Eredmények a *vásárol* kiemelt célszóra (%).

Szerep	**Pontosság**	**Fedés**	**F-mérték**
Vevő	69,88	49,77	**57,63**
Eladó	82,10	60,30	**68,70**
Áru	80,72	77,11	**78,70**
Ár	90,38	83,02	**85,78**
Idő	78,75	52,82	**61,27**
Átlag:	**80,37**	**64,60**	**70,71**

6.3 Eredmények a vevő-centrikus célszavakra

4. táblázat. Eredmények a vevő-centrikus célszavakra (%).

Szerep	Pontosság	Fedés	F-mérték
Vevő	76,01	57,33	**65,09**
Eladó	79,57	66,15	**71,58**
Áru	79,18	80,93	**79,94**
Ár	87,78	80,07	**82,47**
Idő	83,13	63,89	**71,26**
Átlag	**81,13**	**69,67**	**74,07**

A 3. és a 4. táblázat eredményeit összehasonlítva látható, hogy ha a hasonló viselkedésű célszavakat *egy csoportban kezeltük*, akkor majdnem minden esetben jobb eredményeket értünk el, mint ha a célszavakat önállóan vizsgálnánk. Ennek oka, hogy a több célszó több mondatot és jelöltet ad meg és a több jelölt jellemzőiből általánosabb szabályokat tudott készíteni az osztályozó. A modell legjobban az *Ár* és az *Áru* szerepekre, leggyengébben pedig a *Vevő* szerepre teljesített.

6.4 Eredmények az eladó-centrikus célszavakra

5. táblázat. Eredmények az eladó-centrikus célszavakra (%).

Szerep	Pontosság	Fedés	F-mérték
Vevő	74,59	66,82	**70,13**
Eladó	68,97	48,51	**56,35**
Áru	85,92	82,16	**83,64**
Ár	83,64	63,87	**71,58**
Idő	76,38	51,78	**59,86**
Átlag	**77,90**	**62,63**	**68,31**

Az eladó-centrikus esetben a modell legjobban az *Ár* és az *Áru* szerepekre, leggyengébben pedig az *Eladó* szerepre *teljesített*.

6.5 Eredmények a célszavak csoportosítása nélkül

6. táblázat. Eredmények a célszavak csoportosítása nélkül (%).

Szerep	Pontosság	Fedés	F-mérték
Vevő	76,93	60,11	**67,05**
Eladó	72,04	50,39	**59,13**
Áru	83,62	80,24	**82,01**
Ár	88,47	76,26	**81,77**
Idő	85,44	64,53	**73,14**
Átlag	**81,30**	**66,31**	**72,62**

A *célszavak csoportosításától* azt vártuk volna, hogy a Vevő-centrikus célszavaknál a Vevő szerepre, az Eladó-centrikus célszavaknál pedig az Eladó szerepre jobb eredményt kapunk, mint a csoportosítás nélküli esetben. Ez az Eladó szerepre nem teljesült. Ennek egyik oka, hogy az Eladó-centrikus mondatokban az Eladó szerep sokszor távolabb volt a célszótól az elemzőfában. Másik oka, hogy a Vevő-centrikus célszavaknál az Eladó szerepre jó eredményt kaptunk (*71,58-es F-mérték*) a *JeloltVegenBigram* és a *JeloltVegenTrigram* jellemzők hatására. Ez a jó eredmény javította erősen a csoportosítás nélküli esetben is az Eladó szerep eredményét. Így a jobb eredményt a csoportosítás nélküli esetre kaptunk (*72,62-es F-mérték*).

Az eredmények összehasonlítás a kapcsolódó munkákkal. Angol nyelvű szövegekre Gildea és társa [8] sok keretre és azokon belül sok szerepre végezték el a feladatot. Elsősorban igékkel foglalkoznak, de keresnek nem igei célszavakra is. Ezek átlagolt eredményére 63%-os F mértéket kaptak. Eredményeink (*72,62% F-mérték átlag*) jónak számítanak annak ellenére, hogy mi csak egy keretet és ahhoz csak öt főszerepet vizsgáltunk, és csak igei és főnévi igenevekhez kerestünk szerepeket.

7 Összegzés

Munkánkban bemutattunk gazdag jellemzőtéren alapuló gépi tanuló megközelítésünket, amely automatikusan képes magyar nyelvű szövegekben szemantikus szerepek címkézésére függőségi elemző alkalmazásával. A *vállalati vásárlások, tulajdonváltozások* keretével foglalkoztunk. Ezen a kereten belül 1000 annotált mondatot dolgoztunk fel és a következő szerepeket kerestük: *Vevő, Eladó, Áru, Ár, Idő*. *Jellemzőkészletünkben* felszíni, morfológiai és a függőségi elemzés alapján kinyert jellemzőket használtunk fel. Ezen alapjellemzőket kiegészítettük a jellemzőkből számolt *statisztikai arányokkal* is. Megvizsgáltuk, hogy a statisztikai jellemzők hogyan befolyásolják a modell hatékonyságát. Megvizsgáltuk, hogy a modell hogyan teljesít *egy gyakori célszóra önállóan*, és a *célszavak keretekbe összefoglalt csoportjára* is. A mérésekhez célszavainkat csoportosítottuk több szempont szerint. Bár munkánkban a vizsgált szövegek kevesebb témát fedtek le, mint az angol nyelvű szövegekre bemutatott munkák, de eredményeink jónak számítanak a bemutatott angol munkák eredményeivel összehasonlítva.

Hivatkozások

1. Ahn, D.: The stages of event extraction. In: Proceedings of the Workshop on Annotating and Reasoning about Time and Events (ARTE) (2006) 1–8
2. Carreras, X., Màrquez, L.: Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL) (2004) 152–164
3. Carreras, X., Marquez, L., Chrupała, G.: Hierarchical recognition of propositional arguments with perceptrons. In: Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL), Boston, MA (2004) 106–109

4. Carreras, X., Màrquez, L.: Introduction to the CoNLL-2005 shared task: semantic role labeling. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL) (2005) 89–97
5. Ehmann, B., Lendvai, P., Mihăltz, M., Vincze, O., László, J.: Szemantikus szerepek a narratív kategoriális elemzés (NARRCAT) rendszerében. In: Tanács, A., Vincze, V., eds.: IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 121–123
6. Farkas, R., Konczer, K., Szarvas, Gy.: Szemantikus keret illesztés és az IE-rendszer automatikus kiértékelése. In: II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2004) 49–53
7. Fillmore, C.L., Ruppenhofer, J., Baker, C.F.: Framenet and representing the link between semantic and syntactic relations. In: Huang, Ch. and Lenders, W, eds.: Frontiers in Linguistics, volume I of Language and Linguistics Monograph Series B, Institute of Linguistics, Academia Sinica, Taipei (2004) 19–59
8. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics Journal 28/3, (2002) 245–288
9. Johansson, R., Nugues, P.: Semantic structure extraction using nonprojective dependency trees. In: Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval), Prague, Czech Republic (2007) 227–230
10. Koomen, P., Punyakanok, V., Roth, D., Yih, W.: Generalized inference with multiple semantic role labeling systems. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL), Ann Arbor, MI. (2005) 181–184
11. Litkowski, K. C.: SENSEVAL-3 TASK: Automatic Labeling of Semantic Roles. In: Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (2004)
12. Márquez, L., Carreras, X., Litkowski, K.C., Stevenson, S.: Semantic Role Labeling: An Introduction to the Special Issue. Computational Linguistics 34/2 (2009) 145–159.
13. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: An annotated corpus of semantic roles. Computational Linguistics 31/1 (2005) 71–105
14. Pradhan, S., Hacioglu, K., Ward, W., Martin, J.H., Jurafsky, D.: Semantic role chunking combining complementary syntactic views. In: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL), Ann Arbor, MI. (2005) 217–220
15. Subecz, Z.: Detection and Classification of Events in Hungarian Natural Language Texts. Proceedings of the 17th International Conference, TSD 2014, Brno, Czech Republic (2014), Springer Lecture Notes in Computer Science 8655 (2014) 68–75
16. Subecz, Z, Nagyné, Cs.É.: Igei események detektálása és osztályozása magyar nyelvű szövegekben. In: Tanács, A., Varga, V., Vincze, V., eds.: X. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2014) 237–247
17. Surdeanu, M., Marquez, L., Carreras, X., Comas, P.R.: Combination strategies for semantic role labeling. Journal of Artificial Intelligence Research (JAIR) 29 (2007) 105–151
18. Toutanova, K., Haghighi, A., Manning, C.: Joint learning improves semantic role labeling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI (2005) 589–596
19. Xue, N., Palmer, M.: Calibrating features for semantic role labeling. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain (2004) 88–94
20. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In: Tanács, A., Vincze, V., eds.: IX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2013) 368-374